

Judging a Book by its Cover

Brian Kenji Iwana^{*†}, Seiichi Uchida^{*}

^{*}Department of Advanced Information Technology, Kyushu University, Fukuoka, Japan

Email: brian@human.ait.kyushu-u.ac.jp, uchida@ait.kyushu-u.ac.jp

[†]Institute of Decision Science for a Sustainable Society, Kyushu University, Fukuoka, Japan

Abstract—Book covers communicate information to potential readers, but can the same information be learned by computers? We propose a method of using a Convolutional Neural Network (CNN) to predict the genre of a book based on the visual clues provided by its cover. The purpose is to investigate whether relationships between books and their covers can be learned. However, determining the genre of a book is a difficult task because covers can be ambiguous and genres can be overarching. Despite this, we show that a CNN can extract features and learn underlying design rules set by the designer to define a genre. Using machine learning, we can bring the large amount of resources available to the book cover design process.

Index Terms—Convolutional Neural Network; Image Classification; Design; Book Cover

I. INTRODUCTION

“Don’t judge a book by its cover” is a common English idiom meaning not to judge something by its outward appearance. But, when a reader encounters a book, it still happens. The cover of a book is often the first interaction with a book and it creates an impression on the reader and starts the first communication. It starts a conversation with the potential reader and begins to draw a story revealing the contents within. But, what does the book cover say? What are the clues that the book cover reveals? While the visual clues can communicate information to humans, we explore the possibility of using machines to learn about a book by its cover. Machine learning provides the ability to use a large amount of resources and history to the world of design. By bridging the gap between design and machine learning, we hope to use a large data set to understand the secrets of visual design.

We propose a method deriving a relationship between book covers and their genre automatically. The goal is to determine if genre information can be learned based on the visual aspects of a cover created by the designer. This research can aid the design process by revealing underlying information, help the promotion and sales process by providing automatic genre suggestion, and be used in computer vision fields.

The difficulty of this task is that books come with a wide variety of book covers and styles, including nondescript and misleading covers. Unlike other object detection and classification tasks, genres are not concretely defined. Also, a massive amount of books exist and it is not suitable for exhaustive search methods.

To tackle this task, we present the use of an artificial neural network. The concept of neural networks and neural coding is to use interconnected nodes to work together to capture information. Early neural network-like models such as

multilayer perception learning were invented in the 1970s but fell out of favor [1]. More recently, artificial neural networks have been a focus of state-of-the-art research because of their successes in pattern recognition and machine learning. Their successes are in part due to the increase in data availability, increase in processing power, and introduction of GPUs [2].

Convolutional Neural Networks (CNN) [3], in particular, are multilayer neural networks that apply learned convolutional kernels, or filters, to transform input data as a method of feature extraction. The general idea is to use learned features rather than predefined features as the feature representation for image recognition. Recent deep CNNs combine multiple convolutional layers along with fully-connected layers. By increasing the depth of the network, higher level features can be learned and discriminative parts of the images are exaggerated [4]. These deep CNNs have had successes in many fields including digit-recognition [3], [5] and large-scale image recognition [6]–[8].

The contribution of this paper is to demonstrate that connections between book genres can be learned using only the cover images. To solve this task, we implemented a CNN to predict the genre of books based on an image of their cover. We also reveal some of the relationships between genres automatically learned by the CNN. Secondly, we created a large data set for the use of the experiment made of book cover images, title and author text, and category membership. This data set can be used for a variety of tasks some of which include text recognition, font analysis, and genre prediction.

The remaining of this paper is organized as follows. Section II provides related works in learning design with machine learning. Section III elaborates on CNNs and the details of the proposed method. In Section IV, we confirm the proposed method and analyzed the experimental results. Finally, Section V draws the conclusion.

II. RELATED WORKS

Visual design is intentional and serves a purpose. It has a rich history and exploring the purpose has been extensively analyzed by designers [9] but it is a relatively new field in machine learning.

Techniques have been used to identify artistic styles and qualities of paintings and photographs [10]–[13]. Gatys, et al. [11] used deep CNNs to learn and copy the artistic style of paintings. Similarly, the goal of this trial is to learn the stylistic qualities of the work, but we go beyond to learn the underlying meaning behind the style.

In the field of genre classification, there have been attempts to classify music by genre [14]–[16]. It was also done in the fields of paintings [10], [17] and text [18], [19]. Most of these methods use designed features or features specific to the task.

III. CONVOLUTIONAL NEURAL NETWORKS

Modern CNNs are made up of three components: convolutional layers, pooling layers, and fully-connected layers. The convolutional layers consist of feature maps produced by repeatedly applying smaller filters across the input. The filters are weighted and learned in the backpropagation step of neural network training. The resulting feature maps are fed into an activation function after which can be down-sampled by a max pooling [20] layer to reduce the computational time for future layers. Finally, the last few layers of a CNN are typically made up of a few fully-connected layers. These layers are fed a vector representation of the images from a preceding pooling layer and continue like a standard feedforward neural network.

We constructed a CNN with two convolutional layers and two fully-connected layers. The first convolutional layer receives an input of 56px by 56px images with RGB channels. It uses 32 filters of size $5 \times 5 \times 3$, stride 1 and then sampled with max pooling of size 2×2 , stride 1. The second convolutional layer has 64 filters of size $5 \times 5 \times 32$, stride 1 and a max pooling of size 2×2 , stride 1. The results of the second max pooling provide the first fully-connected layer with a vector of length 12,544 ($14 \times 14 \times 64$) which are used by 512 neurons. The final fully-connected output layer uses a 20-wide softmax [21] which represents the probability of each respective 20 class labels. This architecture is similar to the LeNet model [3], but with using rectified linear unit (ReLU) [22] activation functions instead of sigmoid activation functions. We also use dropout [23], a technique to prevent overfitting, with a keep probability of 0.5 for the fully-connected layers.

We used adaptive moment estimation (Adam) [24] for the stochastic optimization with a learning rate of 0.0001. The Adam algorithm updates the moment estimate with parameters based on the estimated moving averages of the gradient and squared gradient with bias estimates. The advantage of using Adam optimization over standard stochastic gradient decent (SDG) is that the learning rates are adaptive and do not require stationary objectives.

IV. EXPERIMENTAL RESULTS

A. Data set preparation

We used a data set consisting of 137,788 unique book cover images classified in 20 categories with a varying but similar distribution of examples. The data set was created from book cover images and genres supplied by Amazon.com [25]. Genres are defined as the top categories under “Books.” The categories are as follows, “Art & Photography,” “Biographies & Memoirs,” “Business & Money,” “Children’s Books,” “Comics & Graphic Novels,” “Computers & Technology,” “Cookbooks, Food & Wine,” “Health, Fitness & Dieting,” “History,” “Humor & Entertainment,” “Law,” “Medical Books,” “Mystery, Thriller & Suspense,” “Politics & Social

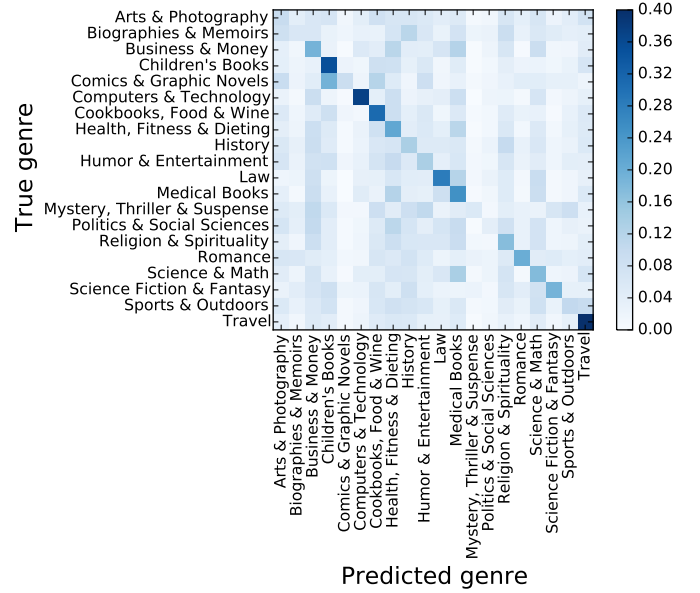


Fig. 1. Confusion matrix for the classification rate accuracy of all categories.

Sciences,” “Religion & Spirituality,” “Romance,” “Science & Math,” “Science Fiction & Fantasy,” “Sports & Outdoors,” and “Travel.” The experiment only considers a single category per book, chosen at random when a book appears in multiple categories. In addition, we resized all of the images to fit 56px by 56px by 3 color channel input of the CNN. The data set also includes title, author, and subcategories for each book, but this information was discarded for the experiment.

B. Evaluation

To demonstrate the classification ability of the proposed method, we randomized and split the data set into three sets at 80% training set, 10% validation set, and 10% test set. No pruning of cover images and no class membership corrections were done. We trained the CNN with a batch size of 500 over 70 epochs. The results of using the training set showed a test set classification accuracy of 21.9% for Top 1, 32.1% for Top 2, and 40.2% for Top 3 which are 4.38, 3.21, and 2.68 times better than random chance, respectively. This shows that classification of book cover designs is possible, although a very difficult task.

A confusion matrix for Top 1 is shown in Fig. 1 with the genre labels used in the experiment. The confusion matrix shows that some classes such as “Computers & Technology” and “Travel” are significantly easier to classify, while “Biographies & Memoirs” and “Politics & Social Sciences” is more difficult. By visualizing the softmax activations, in Fig. 2, we can see an overview of the probability of class membership that the network determined for the book covers. The visualization shows that “Computers & Technology” and “Travel” have a higher accuracy rate because of the higher correlation between their cover image and their class axes. “Computers & Technology,” in particular, has a large cluster



Fig. 2. Visualization of the output layer softmax activations. Each point is a 20-dimensional vector where each dimension is the probability of each output class. For visualization purposes, the points are mapped linearly into 2-dimensional subspace with PCA. The arrows represent the axes of each class. The class ground truth is represented by colors, chosen at random. Sample images with high activations from each class are enlarged.

of same class books at the high end of the “Computers & Technology” axis. Whereas, “Biographies & Memoirs” and “Politics & Social Sciences” has a sporadic relationship between the points and the axes.

C. Analysis

The CNN was able to learn certain high-level features of each category. Figure 3 shows the cover image with the highest probability of being in each class. High-level features were automatically correlated to each category, such as portraits for “Biographies & Memoirs,” food for “Cookbooks, Food & Wine,” and landscapes for “Travel.”

In general, most cover images either have a strong activation toward a single class or are ambiguous and could be part of many classes at once, as shown in Fig. 2. An example of this could be seen in the sample images from the “Cookbooks, Food & Wine” category in Fig. 4. When the cover contained an image of food, the CNN predicted the correct class and with a

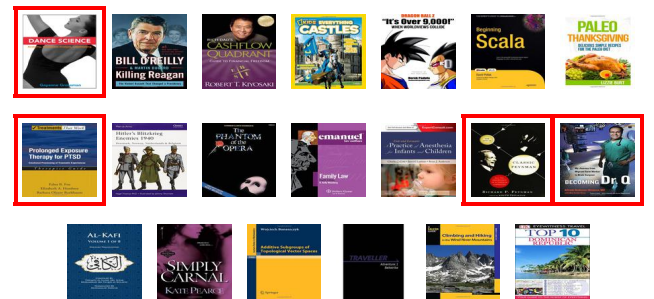


Fig. 3. The test set cover image with the highest softmax activation for each of the 20 classes in the order described in Section IV-A. The images with red borders are misclassifications.

high probability. But, the covers with more ambiguous images resulted in a low confidence. The misclassified examples in Fig. 4 failed for understandable reasons; one features a person



Fig. 4. Sample test set images from the “Cookbooks, Food & Wine” category. The top row shows the cover images and the bottom row shows their respective softmax activations. The blue bar is the correct class and the red bars are the other classes. Only the top 5 highest activations are displayed. The first three images are examples of correctly classified books and the second three are misclassified.

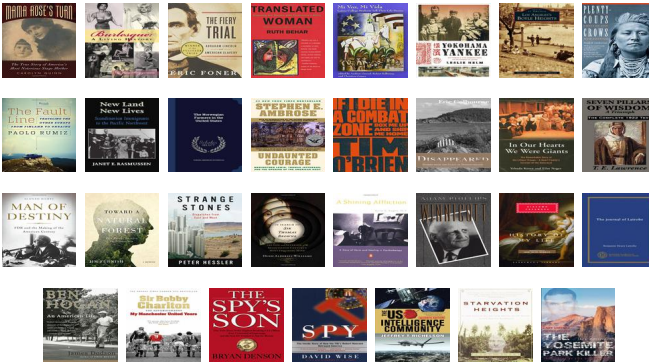


Fig. 5. The “Biographies & Memoirs” book covers that were classified as “History.” While misclassified, many of these books also share “History” as a secondary category.

and was placed in “Romance” or “Biographies & Memoirs” and the next one with a cover image of juice and was placed in “Health, Fitness & Dieting.” The final example has styled cover and was strongly placed in “Arts & Photography.” Misclassifications like this often happens because of a misleading cover image.

Another example of misleading cover images can be seen in the “Biographies & Memoirs” category. The difficulty of this category comes from a high rate of sharing qualities with other categories causing substantial ambiguity of the genre itself. A high number of misclassifications from the “Biographies & Memoirs” category went into “History.” However, Fig. 5 shows that most of those misclassifications could be considered to be part of both categories. We can see a similar relationship between “Comics & Graphic Novels” and “Children’s Books” and between “Medical Books” and “Science & Math.” This shows that the network was able to automatically learn relationships between categories based solely on the cover images.

Misclassifications also happen frequently due to the simplicity of cover designs. In most categories, there exist books with simple solid color designs. But, they are most common in the “Law” and “Religion & Spirituality” categories. This causes many misclassifications of solid color covers to be classified



Fig. 6. (a) shows all of the “Politics & Social Sciences” covers from the test set that were incorrectly classified as “Religion & Spirituality.” (b) shows all of the “Politics & Social Sciences” covers incorrectly classified as “Law.”

as “Law” or “Religion & Spirituality” as learned by the network. It is interesting to note that in absence of distinctive object features, the CNN put weight in color channels to classify covers. This can be seen in Fig. 6 where (a) the solid red “Politics & Social Sciences” covers were classified as “Religion & Spirituality” and (b) “Law” otherwise.

V. CONCLUSION

In this paper, we presented the application of machine learning to predict the genre of a book based on its cover image. We showed that it is possible to draw a relationship between book cover images and genre using automatic recognition. Using a CNN model, we categorized book covers into genres and the results of the experiment had an accuracy of 21.9% for Top 1, 32.1% for Top 2, and 40.2% for Top 3 in 20-class classification.

However, classification of books based on the cover image is a difficult task. We revealed that many books have cover images with few visual features or ambiguous features causing for many incorrect predictions. While uncovering some of the design rules found by the CNN, we found that books can have also misleading covers. In addition, because books can be part of multiple genres, the CNN had a poor Top 1 performance. To overcome this, experiments can be done using multi-label classification.

Future research will be put into further analysis of the characteristics of the classifications and the features determined by the network in an attempt to design a network that is optimized for this task. A larger network may learn more features or design better ones, so increasing the size of the network or tuning the hyperparameters may improve the performance. The book cover data set we created can be used for other tasks as it contains other information such as title, author, and category hierarchy. It can also be used to supplement the CNN with textual features alongside the cover images. We hope to design a network to better capture the essence of cover design.

ACKNOWLEDGMENTS

All book cover images are copyright Amazon.com, Inc. The display of the images are transformative and are used as fair

use for academic purposes.

REFERENCES

- [1] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.
- [2] K. Chellapilla, S. Puri, and P. Simard, "High performance convolutional neural networks for document processing," in *Tenth International Workshop on Frontiers in Handwriting Recognition*. Suvisoft, 2006.
- [3] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [4] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer vision—ECCV 2014*. Springer, 2014, pp. 818–833.
- [5] D. Cireşan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3642–3649.
- [6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.
- [8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [9] J. Drucker and E. McVarish, *Graphic Design History: A Critical Guide*. Pearson Education, 2009.
- [10] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller, "Recognizing image style," *arXiv preprint arXiv:1311.3715*, 2013.
- [11] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015.
- [12] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Computer Vision—ECCV 2006*. Springer, 2006, pp. 288–301.
- [13] —, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys (CSUR)*, vol. 40, no. 2, p. 5, 2008.
- [14] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *Speech and Audio Processing, IEEE transactions on*, vol. 10, no. 5, pp. 293–302, 2002.
- [15] C. McKay and I. Fujinaga, "Automatic genre classification using large high-level musical feature sets," in *ISMIR*, vol. 2004. Citeseer, 2004, pp. 525–530.
- [16] D. Pye, "Content-based methods for the management of digital music," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 6. IEEE, 2000, pp. 2437–2440.
- [17] J. Zujovic, L. Gandy, S. Friedman, B. Pardo, and T. N. Pappas, "Classifying paintings by artistic genre: An analysis of features & classifiers," in *Multimedia Signal Processing, 2009. MMSP'09. IEEE International Workshop on*. IEEE, 2009, pp. 1–5.
- [18] A. Finn and N. Kushmerick, "Learning to classify documents according to genre," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 11, pp. 1506–1518, 2006.
- [19] P. Petrenz and B. Webber, "Stable classification of text genres," *Computational Linguistics*, vol. 37, no. 2, pp. 385–393, 2011.
- [20] D. Scherer, A. Müller, and S. Behnke, "Evaluation of pooling operations in convolutional architectures for object recognition," in *Artificial Neural Networks—ICANN 2010*. Springer, 2010, pp. 92–101.
- [21] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, 2012.
- [22] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–323.
- [23] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [24] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [25] Amazon.com Inc, "Amazon.com: Online shopping for electronics, apparel, computers, books, dvds & more," <http://www.amazon.com/>, accessed: 2015-10-27.